Ecology, 87(10), 2006, pp. 2626–2635

MODEL WEIGHTS AND THE FOUNDATIONS OF MULTIMODEL INFERENCE

WILLIAM A. LINK[1,3] AND RICHARD J. BARKER[2]

[1]*USGS Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, Maryland 20708 USA*
[2]*Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand*

Abstract. Statistical thinking in wildlife biology and ecology has been profoundly influenced by the introduction of AIC (Akaike's information criterion) as a tool for model selection and as a basis for model averaging. In this paper, we advocate the Bayesian paradigm as a broader framework for multimodel inference, one in which model averaging and model selection are naturally linked, and in which the performance of AIC-based tools is naturally evaluated. Prior model weights implicitly associated with the use of AIC are seen to highly favor complex models: in some cases, all but the most highly parameterized models in the model set are virtually ignored a priori. We suggest the usefulness of the weighted BIC (Bayesian information criterion) as a computationally simple alternative to AIC, based on explicit selection of prior model probabilities rather than acceptance of default priors associated with AIC. We note, however, that both procedures are only approximate to the use of exact Bayes factors. We discuss and illustrate technical difficulties associated with Bayes factors, and suggest approaches to avoiding these difficulties in the context of model selection for a logistic regression. Our example highlights the predisposition of AIC weighting to favor complex models and suggests a need for caution in using the BIC for computing approximate posterior model weights.

Key words: AIC; Akaike's information criterion; Bayesian inference; Bayesian information criterion; Bayes factors; BIC; model averaging; model selection; Salmo trutta.

INTRODUCTION

It would be nice if there were no uncertainty about models. In such an ideal world, a single model would be available; the data analyst would be in the enviable position of having only to choose the best method for fitting model parameters based on the available data. The choice would be completely determined by the statistician's theory, a theory which regards the model as an exact depiction of the process that generated the data.

It is completely appropriate and standard practice for statistical methods to be developed in that ideal world; we return to this theme subsequently. But in wildlife and ecological applications, many data sets are observational. Numerous covariates are available, and model selection is an important part of the inferential process. It is clearly wrong to use the data to choose a model and then to conduct subsequent inference as though the selected model were chosen a priori: to do so is to fail to acknowledge the uncertainties present in the model selection process, and to incestuously use the data for two purposes (Chatfield 1995, Draper 1995).

Akaike's information criterion is defined by $\text{AIC} = -2\log(\text{MaxLikelihood}) + 2k$, where k is the number of parameters in the model. Models with smaller values of AIC are favored on the basis of fit and parsimony. AIC weights for a collection of models are proportional to $\exp(-1/2\ \text{AIC})$. For details, see Burnham and Anderson (2002).

The introduction of AIC for model selection and of AIC weights for model averaging have been positive contributions to the fields of wildlife biology and ecology, providing an objective basis for model selection and multimodel inference. The work of Burnham and Anderson (1998, 2002) has been enormously influential in this regard, resulting in a major paradigm shift away from hypothesis testing as a tool for model choice. However, there appears to be a growing resistance to these ideas in the ecological and wildlife literature (e.g., Guthery et al. 2005, Richards 2005, Stephens et al. 2005). Here we offer some thoughts on model selection and model averaging from a Bayesian perspective.

We agree with Burnham and Anderson that it is important to distinguish hypothesis testing (conditional on a model) and the process of selecting the model, and are in agreement with the philosophy that has motivated their work on model selection. That said, we question whether multimodel inference is best accomplished using AIC. Our position is that the Bayesian approach to multimodel inference provides a wider framework in which AIC-based methods can and should be evaluated and alternatives considered.

Our paper is organized as follows: first, we provide an overview of Bayesian multimodel inference, introducing

Manuscript received 14 November 2005; revised 22 March 2006; accepted 4 April 2006. Corresponding Editor: A. M. Ellison.

[3] E-mail: wlink@usgs.gov

notation and basic formulas to be used in the remainder of the paper. Next, we respond to an objection often raised against Bayesian multimodel inference, that "truth in the model set" is an unrealistic and philosophically untenable assumption. We argue that the use of model weights in prediction requires their interpretation as posterior model probabilities. This observation raises the question as to which set of prior model weights is implicitly chosen when one uses AIC weights; the answer provides valuable insights into the operating characteristics of AIC in multimodel inference, explaining its well-documented tendency to favor highly parameterized models (see Kass and Raftery 1995). We recommend that analysts use weighted BIC (Bayesian information criterion) as a computationally simple alternative to AIC, based on explicit selection of prior model probabilities, rather than the default choice implicit to the use of AIC. The weighted BIC (and AIC, as a special case) use approximate rather than exact Bayes factors, which are the fundamental quantities for updating prior to posterior model probabilities. We illustrate difficulties with Bayes factors, and suggest approaches to avoiding them in the context of model selection for a logistic regression. Our example highlights the predisposition of AIC weighting to favor complex models and suggests a need for caution in using the BIC to compute approximate posterior model weights.

An Overview of Bayesian Multimodel Inference

Briefly described, Bayesian multimodel inference (BMI) has three ingredients. The first is a set of models $\mathbf{M} = \{M_1, M_2, \ldots, M_R\}$. Corresponding to model M_i is a probability distribution $f(\boldsymbol{x} \mid \theta^{(i)}, M_i)$ fully specified except for an unknown parameter set $\theta^{(i)}$. It is assumed that one of the models is true, in the sense that the data are a sample from $f(\boldsymbol{x} \mid \theta^{(i)}, M_i)$. The second ingredient for BMI is a set of priors on parameters, one for each model in $\mathbf{M}$; we denote the prior on parameters $\theta^{(i)}$ of model M_i by $g(\theta^{(i)} \mid M_i)$.

The first two ingredients combine to form the marginal distribution

$$\mathrm{P}(\boldsymbol{x}|M_i) = \int f(\boldsymbol{x}|\theta^{(i)}, M_i) g(\theta^{(i)}|M_i) d\theta^{(i)} \qquad (1)$$

which is the average probability distribution under model M_i averaged against the priors for the parameters. Regarded as a function of the model, for fixed data, it serves as a likelihood function for the model. The Bayes factor for comparing models i and j is the ratio of these model likelihoods, namely,

$$\mathrm{BF}_{i,j} = \frac{\mathrm{P}(\mathrm{Data}|M_i)}{\mathrm{P}(\mathrm{Data}|M_j)}.$$

The final ingredient for BMI is a collection of prior probabilities $\{\pi_1, \pi_2, \ldots, \pi_R\}$ assigned to the collection $\mathbf{M}$, independent of the data; $\pi_i = \Pr(M_i)$ is the prior probability that model M_i is true. Bayes' theorem relates posterior to prior model probabilities via the formula

$$\Pr(M_i|\mathrm{Data}) = \frac{\mathrm{P}(\mathrm{Data}|M_i)\Pr(M_i)}{\sum_j \mathrm{P}(\mathrm{Data}|M_j)\Pr(M_j)}. \qquad (2)$$

The Bayes factor for comparing models i and j can be shown to be the ratio of posterior to prior odds, i.e.,

$$\mathrm{BF}_{i,j} = \frac{\Pr(M_i|\mathrm{Data})/\Pr(M_j|\mathrm{Data})}{\Pr(M_i)/\Pr(M_j)}.$$

Dividing numerator and denominator on the right-hand side of Eq. 2 by $\mathrm{P}(\mathrm{Data}|M_1)$, we obtain

$$\Pr(M_i|\mathrm{Data}) = \frac{\mathrm{BF}_{i,1}\pi_i}{\sum_j \mathrm{BF}_{j,1}\pi_j}. \qquad (3)$$

Thus, it is seen that Bayes factors provide a mechanism for converting prior model probabilities to posterior model probabilities. Posterior model probabilities are used both for model selection and model averaging: if we wish to identify the best supported models in $\mathbf{M}$, the choice is naturally made on the basis of these posterior probabilities; if we wish to produce a model-averaged prediction, the laws of probability lead to

$$\Pr(\mathrm{Prediction}|\mathrm{Data}) = \sum_i \Pr(\mathrm{Prediction}|\mathrm{Data}, M_i)\Pr(M_i|\mathrm{Data}). \qquad (4)$$

Thus, model selection and model averaging are naturally linked under BMI. For a thorough introduction to Bayesian multimodel inference, refer to Draper (1995), Hoeting et al. (1999), and Wintle et al. (2003). Frequentist alternatives are considered in Hjort and Claeskens (2003) and Claeskens and Hjort (2003); in the subsequent discussion, Raftery and Zheng (2003) argue convincingly for the superiority of Bayesian methods, even when evaluated from a Frequentist perspective.

One final piece of background material will be useful for our discussion. The Bayesian information criterion is defined by

$$\mathrm{BIC}_i = -2\log[f(\mathrm{Data}|\hat{\theta}^{(i)}, M_i)] + k_i \log(n)$$

where $\hat{\theta}^{(i)}$ is the maximum likelihood estimator of the parameters for model i, k_i is the number of parameters in model i, and n is the sample size. This quantity can be used to construct an asymptotic approximation to $\mathrm{BF}_{i,j}$, namely $\exp(-(\mathrm{BIC}_i - \mathrm{BIC}_j)/2$ (Kass and Raftery 1995). Substituted in Eq. 3, we can obtain approximate posterior probabilities:

$$\Pr(M_i|\mathrm{Data}) \approx \frac{\exp(-\mathrm{BIC}_i/2)\pi_i}{\sum_j \exp(-\mathrm{BIC}_j/2)\pi_j}. \qquad (5)$$

Assigning uniform prior probabilities to the set $\mathbf{M}$, $\pi_i \equiv 1/R$, yields what are commonly referred to as BIC weights; Eq. 5 can be thought of as a generalized BIC weight.

A Preliminary Objection Addressed

It is sometimes claimed that Bayesian multimodel inference (BMI) is philosophically unsatisfactory on the grounds that it requires that "truth be in the model set." We believe this objection to be overstated.

Some might object to there even being such a thing as "Truth"; others might be willing to concede that such a thing could exist, but that we would never be able to identify it if we came across it. Most would concede that it is unlikely that Truth is in our model set. But this discussion, while entertaining philosophically, is a red herring with regard to the utility of Bayesian model averaging. *Conditioning on ''Truth in the model set'' is no less innocent than conditioning on individual models for the purpose of developing estimators in parametric analysis*. Maximum likelihood estimators used in calculating AIC values are derived assuming the individual models are true, regardless of whether one believes the model to be a perfect depiction of data generating mechanisms; the estimates are interpreted conditionally, that is to say, in context of this assumption. Similarly, Bayesian model selection and model averaging are conducted and interpreted having conditioned on a model set, without requiring intellectual certainty that any one of the models is Truth.

Let us concede for the moment that such a thing as Truth exists, and is one of some vast collection of potential models $\mathcal{M}$, of which $\mathbf{M}$, as previously described, is only a small subset. We may conceive of Nature's choice as a multinomial experiment, a single value drawn from $\mathcal{M}$. It could be that Truth is in $\mathbf{M}$; it could be otherwise. One way or the other, we may carry out the Bayesian calculus conditioning on the event that nature's draw from $\mathcal{M}$ was in $\mathbf{M}$; that is, that truth is in our model set. The subsequent process of updating prior probabilities to posterior probabilities is unchanged.

But doesn't this mean that we are basing our analysis on a hypothetical we presume to be false? Of course it does! We do similar things all the time in scientific endeavors. Box (1976) said it well:

> The statistician knows, for example, that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world.

Model-based statistical inference uses methods developed under assumptions that the models considered are "true." Inferentially valid application of these methods does not require that the real world conform exactly with the model, but that the model be a good approximation. The mathematician doesn't question whether X is really a normal random variable in developing the methods, or even whether such a thing exists in the real world. Application of the methods is conditional on the assumption, which is not the same as saying we believe it is "truth," but only that it is "close enough to truth" so as not to misguide our decisions.

So then, rather than say that Bayesian multimodel inference "requires that truth be in the model set" we would say Bayesian multimodel inference *operates as though truth were in the model set*. Berger and Pericchi (1996) refer to "truth in the model set" as "standard Bayesian language" and note that "one does not strictly have to assume that one of the models is true." They suggest that Bayes factors "be interpreted solely in terms of comparative support of the data" for the various models. Bayesian multimodel inference uses "truth in the model set" as a model itself, rather than as a statement of reality.

We suggest that AIC model averaging implicitly uses the same structure as BMI, conditioning on "truth in the model set." This becomes clear once one considers model weights as model probabilities.

Model Weights As Model Probabilities

Model weights have the property of being nonnegative and summing to one. The weight on a collection of models (say, all those containing a specific parameter of interest) is obtained by adding up the weights of the individual models in the collection. For a finite space of outcomes, these are the defining characteristics of a probability measure. Model weights *are* probabilities. But probabilities of what?

Burnham and Anderson (2004:272) describe the model weight w_i calculated using AIC as the probability of the event "that model i is, in fact, the K-L best model for the data." Here, K-L refers to Kullback-Leibler distance; the "K-L best model" is the one in the model set closest to Truth. Given that AIC is an estimator of K-L distance, the interpretation that Burnham and Anderson suggest cannot be supported. Suppose that circumstances were such that AIC approximated K-L distance to a high degree of accuracy, so that there could be no uncertainty in the model rankings, and suppose that there were a unique K-L best model in the model set. The minimum AIC model, then, would have to be the K-L best model, even though its AIC weight need not be 100%. Thus, AIC weights cannot be interpreted as probabilities "that model i is, in fact, the K-L best model for the data."

The simplest explanation of model weight w_i is as the probability that M_i is truth, given that truth is in the model set $\mathbf{M}$; we shall argue that this is not only the simplest interpretation, but also the *only* mathematically legitimate interpretation. No philosophical complications are attached to it, just as no philosophical complications are attached to using model-specific likelihoods to compute estimates and standard errors. Estimates for model i are obtained assuming that M_i is Truth, not K-L best.

Readers might ask whether there really is a practical difference in interpretations of model weights. We will return to this point presently, after first establishing our

claim that the simple explanation of model weights offered in the previous paragraph is the only possible interpretation.

Consider the use of model weights given by Burnham and Anderson (2002:448): "... if a parameter θ is common over all models (as θ_i in model M_i), or our goal is prediction, by using the weighted average, we are basing point inference on the entire set of models,

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i \ldots." \quad (6)$$

Predictions are based on conditional distributions of unobserved quantities, given present data and given one or more models. The basis of model-averaged predictions is the average of the conditional distributions, namely,

$$\Pr(\text{Prediction}|\text{Data}) = \sum_{i=1}^{R} w_i \Pr(\text{Prediction}|\text{Data}, M_i). \quad (7)$$

Note that this is of the same form as Eq. 6, though there are no circumflexes indicating the use of estimated parameters in the calculation. If the parameters θ_i were known without error (or nearly so), the circumflexes could also be dropped in Eq. 6; there is no formal accounting for the uncertainty in parameter estimation in the definition of AIC weights.

Subject only to mild assumptions, it can be shown that Eq. 7 is true if and only if

$$w_i = \Pr(M_i|\text{Data}).$$

(The requisite condition is that $M_i \neq M_j$ implies $\Pr(\text{Data} \mid M_i) \neq \Pr(\text{Data} \mid M_j)$, i.e., that there are no identical models, in terms of probability, in the model set.) Thus, w_i must be interpreted as the probability that model M_i is true, given that truth is in the model set. Burnham and Anderson note "an interesting and recent finding...that AIC can be derived under a formal Bayesian framework." We take the point a step further, saying that model weighting has no compelling epistemological foundation outside of the Bayesian paradigm.

Why it matters

Our point in this and the preceding section is that the issue of whether "truth is in the model set" ought to be laid aside as an irrelevancy in comparing approaches to model weighting. Focus instead should be on what is assumed in the model(s) and the consequences of prior model weights, whether these are chosen explicitly or implicitly; we explore the latter possibility subsequently. We reiterate that Bayesian multimodel inference uses "truth in the model set" as a model itself, rather than as a statement of reality. In this context "the probability of a model" is always conditional on a model set, and can be interpreted as a relative degree of support within that set. What is more, seeing model weights as model probabilities provides a natural link between model selection and model averaging: models are selected and weighted on the basis of high probabilities.

Once one is willing to regard model weights as probabilities, the full benefits of the calculus of probabilities can be brought to bear on the objects of inference. In particular, Eqs. 2 and 3 can be used to relate model weights to prior probabilities.

Bayesian inference typically begins with the specification of prior probabilities, combining these with probabilities for observed quantities to produce posterior probabilities via Bayes' theorem. On the other hand, we may use Eqs. 2 and 3 to work backwards from posterior probabilities to prior probabilities. It is then possible to evaluate a set of model weights in terms of implicit prior weights, asking what prior weighting scheme leads to this set of weights as posterior model weights.

Burnham and Anderson (2004) have essentially done this, though using the approximation (Eq. 5) instead of the exact formula (Eq. 3). Substituting

$$\pi_i = \frac{\exp[k_i \log(n)/2 - k_i]}{\sum_{r=1}^{R} \exp[k_r \log(n)/2 - k_r]} \quad (8)$$

in Eq. 5, one obtains

$$\Pr(M_i|\text{Data}) \approx \frac{\exp\left(-\frac{1}{2}\text{AIC}_i\right)}{\sum_{r=1}^{R} \exp\left(-\frac{1}{2}\text{AIC}_r\right)},$$

(the AIC weight). Thus, they named weights π_i, defined by Eq. 8, the "K-L [Kullback-Leibler] prior"; this prior distribution leads to AIC weights as approximate posterior model probabilities.

It is worth noting that the K-L prior does not depend on the data, as might be expected in calculating an implicitly defined prior from a set of model weights. It does, however, depend on the sample size. For fixed $n > 7$, it is clear that the larger the value of k_j, the larger the value of $\exp(k_j \log(n)/2 - k_j)$; if $k_j > k_h$, model j will be preferred a priori over model h. The difference in prior weight can be surprisingly large, as will be seen in our subsequent example. Given Burnham and Anderson's perspective that truth is infinite dimensional (Burnham and Anderson 1998:11), priors similarly depending on n and k have some appeal: Burnham and Anderson label such as "savvy priors."

Such priors are unconventional in Bayesian analysis. An appealing feature of Bayesian analysis is that in all but pathological cases, posterior inference is increasingly influenced by the data, rather than the prior, as sample size increases. It is said that the data "overwhelm the prior." If, however, the prior is allowed to vary as sample size increases, the data may not overwhelm the prior. An illustration of the pathological effects of allowing the prior to depend on the sample size is given in the Appendix: the posterior mean can be an inconsistent estimator, i.e., one which converges in

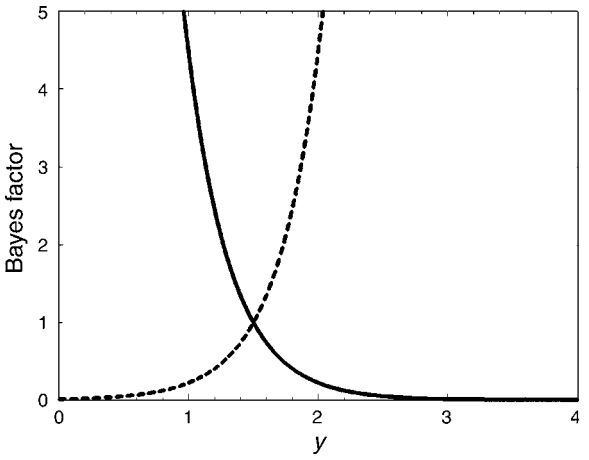

FIG. 1. Bayes factors $BF_{1,2}(y)$ (solid line) and $BF_{2,1}(y)$ (dashed line). $BF_{i,j}(y)$ is the relative support for model i vs. model j based on observation y.

probability to the wrong value as sample size increases without bound.

One must ask whether a "savvy" prior will not produce similarly undesirable results in multimodel analysis. Our point is not to dismiss AIC weights, but to argue the importance of knowing what one's methods are doing.

Part of the appeal of AIC lies in the simplicity of its calculation. The generalized BIC weights (Eq. 5) are equally easy to calculate, but allow the specification of prior model weights, rather than passive acceptance of a default prior. If the default prior is used, it is important that the implications of this choice be fully understood; if the default prior does not reasonably summarize prior beliefs, then other priors must be considered. As stressed elsewhere (Anderson et al. 2001), a crucial component of reporting Bayesian analyses is reporting the priors used; it is, in our view, a mistake to unquestioningly use the K-L prior.

Calculation of approximate posterior model weights using Eq. 5 is easy. Unfortunately, there are serious problems to be confronted in using Bayesian multimodel inference. It is tempting to sweep these under the rug, especially for practitioners likely to be boggled by what appear to be mathematical niceties. However, failure to recognize the subtle relationships between posteriors and priors inherent in multimodel inference can have profound implications. Model selection and model averaging are deep waters, mathematically, and no consensus has emerged in the substantial literature on a single approach. Indeed, our only criticism of the wide use of AIC weights in wildlife and ecological statistics is with their uncritical acceptance and the view that this challenging problem has been simply resolved.

The difficulties associated with Bayesian multimodel inference result from selection of priors for parameters and their effects on Bayes factors. We illustrate these problems in the next section, then suggest a solution in the subsequent section in the context of an example.

Problems with Bayes Factors

We begin by considering two simple models, neither containing unknown parameters; in this case, the marginal distribution is simply $f(\text{Data} \mid M_i)$. Suppose that the data is an observation, Y, which is either sampled from a standard normal distribution ($M = 1$) or from a normal distribution with mean 3 and variance 1 ($M = 2$). The Bayes factor based on an observation $Y = y$ is

$$BF_{1,2}(y) = \frac{f(y|M_1)}{f(y|M_2)} = \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)}{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(y-3)^2\right]} = \exp\left(\frac{9}{2} - 3y\right). \quad (9)$$

Not surprisingly, this is a rapidly decreasing function of y, equal to one if and only if $y = 1.5$ halfway between the two hypothesized means (Fig. 1). For $y = 1$, the Bayes factor $BF_{1,2}(y)$ is approximately 4.48: we might say that the evidence favors model 1 over model 2 by a factor of 4.48 to 1. Eq. 3 provides the means for evaluating this evidence in light of prior knowledge.

Problems with the Bayes factors relate to the expression of uncertainty in model parameters and are most in evidence when alternative models have varying numbers of unknown parameters. To illustrate, suppose that in the foregoing example we retain from Model 1 that Y is standard normal, but modify Model 2 to be that the mean is somewhere in the neighborhood of 3, but not exactly at 3. For example, we might specify Model 2 as implying that $f(y \mid \mu, M_2) = N(\mu, 1)$, with our uncertainty about μ expressed by the prior $g(\mu \mid M_2) = N(3, \sigma^2)$; here, and subsequently the notation $N(m, v)$, denotes a normal distribution with mean m and variance

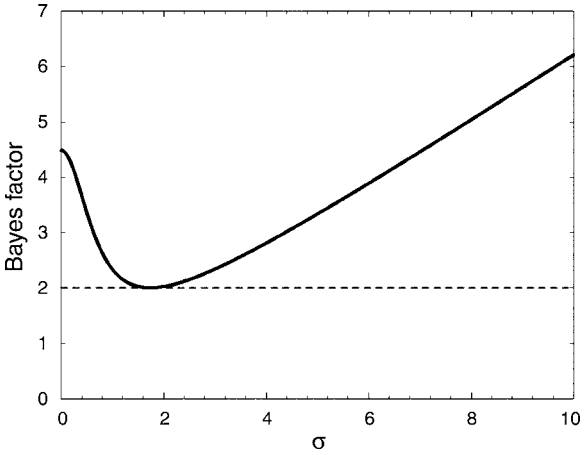

FIG. 2. Bayes factor $BF_{1,2}(1)$ (the relative support for model 1 vs. model 2, based on an observation $y = 1$) as function of uncertainty σ under model 2.

v. Using Eq. 1, it follows that $f(y \mid M_2) = N(3, 1 + \sigma^2)$, so the Bayes factor is

$$\mathrm{BF}_{1,2}(y) = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)}{\frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\left[-\frac{1}{2(1+\sigma^2)}(y-3)^2\right]} = \sqrt{1+\sigma^2} \exp\left[-\frac{\sigma^2 y^2 + 6y - 9}{2(1+\sigma^2)}\right].$$

It is interesting to examine the effect of the prior uncertainty on the Bayes factor. Fig. 2 displays the Bayes factor for an observation $y = 1$, as a function of σ. Note that the evidence in favor of Model 1 starts at the value 4.48 previously calculated, when $\sigma = 0$, then drops to a minimum of 2.0 when $\sigma = 1.73$ ($\sigma^2 = 3$). The performance of the Bayes factor is reasonable: as σ increases from zero, the plausibility under Model 2 of an observation $y = 1$ increases. With $\sigma^2 = 3$, the marginal variance of Y under Model 2 is 4; hence, $y = 1$ is only one standard deviation away from the mean, as opposed to 2 standard deviations away when $\sigma^2 = 0$ (recall that the marginal variance is $1 + \sigma^2$). However, as σ^2 increases beyond 3, the marginal distribution under Model 2 becomes increasingly diffuse, making any particular observation increasingly implausible.

For a fixed value of σ, $\mathrm{BF}_{1,2}(y)$ tends toward zero as y tends toward plus or minus infinity: extreme values of y are simply inconsistent with Model 1, which specifies a standard normal distribution for Y. But it is troubling that for a fixed value of y, as σ increases without bound, so does the Bayes factor. Large values of y augur against Model 1, but how large they must be depends on σ. We may be tempted to set σ to a large value in expressing our uncertainty about μ under Model 2, but in so doing we make $\mathrm{BF}_{1,2}(y) \gg 1$ for all but extremely large values of y, thus favoring Model 1 in our analysis.

This observation presents a difficulty for objective Bayesian analysis, in which vague, even improper, priors are placed on parameters. In estimating the mean of a normal distribution, it is a common expedient to treat the mean as having been sampled from a (conjugate) normal prior with infinitely large variance. This expedient is innocuous enough for the purposes of estimation, but calamitous for multimodel inference and model selection: Bayes factors are unstable in the presence of improper, noninformative priors for model parameters, and especially so when there are varying numbers of parameters in different models under consideration. Berger and Pericchi (1998) note that these problems extend to the use of vague proper priors. Loosely speaking, we may identify the problem as that models having more parameters allow greater prior uncertainty in the range of the data to be produced; this is reflected in typically lower values for the marginal distribution function of the data, hence a tendency for the Bayes factor to be large in comparing a simple model to a more complex model. The greater the uncertainty in the collection of priors, the more serious the problem becomes.

The example presented may seem artificial, being based on a sample of size one. However, the small sample size was chosen merely for ease of presentation, and is not the source of the difficulty. Given a sample of n, independent observations of a normal random variable and the same models as before, the Bayes factor can be shown to be

$$\mathrm{BF}_{1,2}(\bar{y}) = \sqrt{1+n\sigma^2} \exp\left\{-3n\bar{y} + \frac{9n}{2} - \left[\frac{n^2\sigma^2}{2(n\sigma^2+1)}\right](\bar{y}-3)^2\right\}$$

where $\bar{y}$ is the sample mean. For fixed n and $\bar{y}$, the quantity in the exponential approximates $-n\bar{y}^2/2$ as σ^2 increases, but the quantity under the radical increases without bound. Thus, the selection of a vague prior for μ inevitably leads to favoring Model 1 over Model 2. Note that conditional on Model 2, the posterior distribution for μ is normal with mean and variance converging to $\bar{y}$ and $1/n$ as σ^2 increases without bound, a perfectly reasonable basis for inference, conditional on the model. The vague prior on μ is harmless for estimation, but has undesirable consequences for model selection.

There has been considerable theoretical work put to the problem of defining stable Bayes factors, primarily through the specification of reasonable default priors for model parameters (Berger and Pericchi 1996, 1998, Kadane and Lazar 2004). The issues are highly technical, and difficult. Nevertheless, the Bayesian paradigm for multimodel inference is quite simple, having two components. First, prior model weights are chosen independent of data. Then, priors for parameters are selected for each model; given data, these allow computation of Bayes factors. Bayes factors are then combined with prior model weights to compute posterior model weights. While there are technical difficulties to be overcome, we suggest that they are not insurmountable and that Bayesian multimodel inference is a philosophically satisfying and self-consistent approach to dealing with model uncertainty; indeed, it is our conviction that there is no valid epistemological basis for model weighting outside of the Bayesian paradigm. Given reasonable choices of priors for parameters, the Bayes factor can be calculated and used as the basis of passing from prior to posterior model weights. Assessing the reasonableness of selected priors is inevitably a subjective process, subjective but honest, if the process of selection is made transparent in the presentation of results. It is, to our mind, far better to lay subjective choices out on the table and to present a mathematically precise analysis, than to ignore automatic choices in approximate analyses and to mistake arbitrariness for objectivity.

Example

Our example is from a study of brown trout (*Salmo trutta*) spawning in a tributary of Lake Brunner, located in the West Coast region, South Island, New Zealand.

TABLE 1. Bayes factors BF(1, j) for comparing models 1 and j, calculated using three different priors on parameters (π_U, π_D, and π_V); comparisons for models i and j can be based on BF (i, j) = BF (1, j) / BF (1, i). The BIC (Bayesian information criterion) approximation for comparing models 1 and j is exp[−(1/2)BIC$_1$]/exp[−(1/2)BIC$_j$].

Prior and BIC	BF (1,2)	BF (1,3)	BF (1,4)	BF(1,5)
π_U	12.4	31.7	281.7	390.1
π_D	13.1	33.8	288.9	355.8
π_V	109.2	282.4	31 603.8	563 509.7
BIC	16.1	42.4	712.3	3723.7

Logistic regression analysis was used to model the return rate of spawning trout one year after they were caught and tagged in the spawning run of June 1987. Interest was in whether the return rate differed between males and females and whether it differed according to the initial length of the fish.

Priors chosen for parameters

All of the models we consider are of the form

$$\eta_i \equiv \text{logit}(p_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}. \quad (10)$$

We consider five models: Model 1: Constant; Model 2: Length; Model 3: Sex; Model 4: Sex + Length; and Model 5: Sex + Length + Sex × Length.

We must choose model-specific priors on the regression coefficients β_j, and in so doing, bear in mind the difficulties inherent to Bayes factors when vague priors are assigned. The example presented shows that these problems relate to varying degrees of total prior uncertainty in the various models. Hence, we propose defining a parameter V, common to all models, representing the total variability of the linear predictor η in Eq. 10. Given that the regressors $x_{j,i}$ have been standardized across i, the total variability of η_i can be shown to be equal to the sum of the prior variances of the β_js.

Suppose that model i has k regressors, hence $k_i + 1$ parameters. Our approach is to assign mean zero normal priors with variance $V/(k_i + 1)$ to the β's included in the model. This way, regardless of the number of parameters in the model, the total prior uncertainty in the linear predictor is fixed.

We must also choose a prior for V. This parameter, acting like a variance, is naturally endowed with an inverse Gamma prior. We considered two choices of prior. In the first, we supposed that $1/V$ has mean $p/\lambda = 3.2890/7.8014 = 0.4216$, and variance/mean ratio of $1/\lambda = 1/7.8014 = 0.1282$. In the second, we supposed that $1/V$ has mean $p/\lambda = 0.001/0.001 = 1.000$, and variance/mean ratio of $1/\lambda = 1/0.001 = 1000$. The latter is a standard vague prior for a variance parameter; the former has the appealing property that if, given V, logit(p) is a mean-zero normal random variable with variance V, then marginally p has a distribution that is approximately uniform on the unit interval. Thus, both priors were chosen to represent absence of prior knowledge of parameters.

For ease of description, we refer to the two priors as π_U (the "U" calling to mind the uniform distribution) and π_D (the "D" calling to mind the more diffuse distribution). Finally, in order to illustrate the problems with Bayes factors resulting from vague priors, we consider a prior set π_V, with independent, identically distributed normal priors on all coefficients in Eq. 10, these having mean of zero and variance equal to 1000. In light of the example in the section on difficulties with Bayes factors and vague priors, we anticipated that π_V would lead to Bayes factors highly and unrealistically favoring the most simple model. We also anticipated that the priors π_U and π_D were sufficiently uninformative to yield posterior distributions for parameters similar to those arising from the use of π_V.

Given a prior on the parameters, the task is to compute Bayes factors. We performed this calculation using Markov chain Monte Carlo (MCMC). The results we report were obtained using Reversible Jump MCMC (RJMCMC; Green 1995), implemented in program GAUSS (Aptech Systems, Black Diamond, Washington, USA). These calculations can also be performed using program WinBUGS (Spiegelhalter et al. 2000) with code available online (see Supplement) though at considerably greater expense in computation time. Having obtained Bayes factors, one may convert priors on models to posterior model weights by means of Eq. 3. We also computed maximum likelihood estimators, and the BIC, in order to approximate the posterior model probabilities by the weighted BIC given in Eq. 5.

Priors on models

We considered four sets of priors for models. For the sake of comparison, we chose the Burnham and Anderson K-L prior. We also considered uniform prior model weights (weight = 1/5 on each of the five models considered); Ockham weights (favoring parsimonious models, with prior weights proportional to exp[−number of parameters]), and Complexity weights (moderately favoring more complex models as a reflection of the notion that Truth is complex, proportional to exp(number of parameters)). Note that for each specification of priors on model parameters, the transformation of prior model weights to posterior model weights involves the same set of Bayes factors.

Computation of Bayes factors

Implementation of Bayesian multimodel inference using MCMC treats "Model" as a latent categorical variable. For each set of priors on parameters, we began by performing an analysis with uniform prior probabilities on models, using the Markov chain output to compute approximate posterior model probabilities; we used the uniform priors and approximate posterior model probabilities to make an initial approximation to the Bayes factors. In the interest of having all five

TABLE 2 Four priors: Ockham [favors parsimony, prior weights proportional to exp(–k)], Constant [equal weights], Complexity [favors complex models, prior weights proportional to exp(k)], and K-L [weights given by Eq. 8], on five models (Constant, Length, Sex, Sex + Length, and Sex × Length).

Prior	Constant	Length	Sex	S + L	S × L
Ockham	0.521	0.192	0.192	0.070	0.026
π_U	0.960	0.028	0.011	0.000	0.000
π_D	0.962	0.027	0.010	0.000	0.000
π_V	0.995	0.003	0.001	0.000	0.000
BIC	0.969	0.022	0.008	0.000	0.000
Constant	0.200	0.200	0.200	0.200	0.200
π_U	0.894	0.072	0.028	0.003	0.002
π_D	0.899	0.069	0.027	0.003	0.003
π_V	0.987	0.009	0.003	0.000	0.000
BIC	0.920	0.057	0.022	0.001	0.000
Complexity	0.029	0.080	0.080	0.218	0.592
π_U	0.723	0.158	0.062	0.019	0.037
π_D	0.729	0.152	0.059	0.019	0.041
π_V	0.966	0.024	0.009	0.000	0.000
BIC	0.801	0.135	0.051	0.008	0.004
K-L	0.0002	0.0035	0.0035	0.0574	0.9353
π_U	0.063	0.088	0.035	0.064	0.751
π_D	0.059	0.079	0.031	0.058	0.773
π_V	0.807	0.129	0.050	0.007	0.007
BIC	0.255	0.257	0.098	0.095	0.296

Note: Beneath each model prior is the set of posterior model weights computed using the Bayes factor determined by π_U, π_D, and π_V, and by the BIC.

models adequately sampled and the Markov chain adequately mixed, we used these approximate Bayes factors to choose model priors that would induce nearly constant posterior model probabilities, then re-ran our analyses generating chains of length 5 000 000 after a burn-in of length 100 000. These specified priors, and the resulting approximate posterior model weights were then used to recalculate the Bayes factors (this method of tuning the MCMC algorithm was suggested by Carlin and Chib [1995]). These calculations took approximately 6.5 hours when implemented using RJMCMC in GAUSS. The simulations were long enough to ensure good mixing of the chains, as indicated by examination of within-chain autocorrelation and comparison of parallel chains; we estimate that the Bayes factors are correct up to a factor of ±2%. We note that run time for our WinBUGS code is about six times longer for chains of the same length. The results we present are nearly identical to results obtained in WinBUGS using chains of length 1 000 000.

Results

Bayes factors and the BIC approximation are given in Table 1. The first thing to note is that, for all of the priors considered, Model 1 is favored over the others; Kass and Raftery (1995) describe weights of evidence in favor of one model over another as Positive ($3 < \text{BF} \leq 20$), Strong ($20 < \text{BF} \leq 150$), and Very Strong (BF > 150). The vague prior π_V, having greater prior uncertainty in more complex models, massively overstates the evidence in favor of Model 1 against the others; had we used prior variances of 100 000 rather than 1000, the overstatement would have been even greater. Priors π_V and π_D, based on partitioning the total prior variance of the linear predictor, so as to avoid the problems evident in analysis based on π_V, yield similar inferences.

Bayes factors are determined by the set of models, the priors chosen for their parameters, and the data, but do not depend on the set of prior model weights. Table 2 presents the set of four priors for the set of five models, and the resulting posterior distributions. We regard the first three priors (Ockham, Constant, and Complexity) as reflecting reasonable levels of prior uncertainty, with the Ockham and Complexity priors representing moderate predispositions toward parsimony and complexity. The K-L prior, however, puts prior weight of >99% in favor of the two most complex models, and odds of 4621:1 against the simplest model. The posterior weights using BIC and the K-L prior are the AIC weights.

In contrast to the K-L prior, each of the first three priors (Ockham, Constant, and Complexity) result in most (>72%) of the posterior model weight being placed on the constant model. It is remarkable that, despite the overwhelming prejudice the K-L prior exhibits against the simplest model, the AIC weights still place comparable weight on the simplest (25.5%) and most complex (29.6%) models, and none of the models appears unreasonable on the basis of posterior probability; the data appear to have fought back against a highly prejudicial prior. However, we are more inclined to trust analyses based on the fully Bayesian analyses with priors π_U and π_D; these priors on parameters do not inequitably influence the Bayes factor comparisons of models (as does π_V), nor are they based on doubtful approximations, as BIC. That said, the K-L prior leads to posterior model weights that still reflect the K-L prior's prejudices: the posterior odds ratio for the most

TABLE 3. Features of posterior distributions for parameters of Model 5 (β_0 = constant, β_1 = length effect, β_2 = sex effect, β_3 = length × sex interaction), resulting from three priors on parameters (π_U, π_D, and π_V).

Prior	Parameter	Mean	SD	2.5%	Median	97.5%
π_D	β_0	−3.03	0.11	−3.25	−3.02	−2.82
	β_1	0.51	0.21	0.10	0.51	0.92
	β_2	0.03	0.11	−0.18	0.03	0.24
	β_3	−0.41	0.21	−0.82	−0.41	−0.01
π_U	β_0	−3.01	0.11	−3.23	−3.01	−2.80
	β_1	0.49	0.21	0.09	0.49	0.89
	β_2	0.03	0.11	−0.18	0.02	0.24
	β_3	−0.40	0.20	−0.79	−0.39	0.00
π_V	β_0	−3.04	0.11	−3.26	−3.04	−2.83
	β_1	0.53	0.21	0.11	0.53	0.95
	β_2	0.03	0.11	−0.18	0.03	0.25
	β_3	−0.43	0.21	−0.84	−0.43	−0.02

Note: Posterior summaries are the mean, standard deviation (SD), 2.5th, 50th (median), and 97.5th percentiles.

complex against the simplest model are 46.4 and 63.5, for π_U and π_D.

Finally, we note that the informative prior, π_U, and the diffuse constant variability prior, π_D, lead to posterior distributions for parameters that are essentially equivalent to those obtained using the vague prior π_V (Table 3). Many Bayesian practitioners would naturally feel more comfortable using π_V as the basis of an objective Bayesian analysis; the problematic effects of vague priors on Bayes factors are avoided using π_U or π_D, without substantial effects on the estimates of model parameters.

Summary and Conclusions

The introduction of AIC weights to the field of wildlife and ecological applications has been a significant and positive development for multimodel inference. However, there is substantial room for improvement; specifically, there is a need for greater understanding of latent assumptions, a need for greater flexibility in implementation, and a need for improved accounting for parameter uncertainty in choosing model weights.

Model averaging has a sound logical foundation in Bayesian inference; in this context, one may examine the tendency of AIC weights to favor complex models, one may choose alternative prior weights reflecting other intellectual predispositions, and one may provide a formal accounting for parameter uncertainty.

Bayesian multimodel inference requires the explicit specification of priors for parameters and priors for models. The set of Bayes factors characterizes all that the models and priors for parameters say about the data; Bayes factors are invariant to the choice of priors on models.

Bayes factors can be sensitive to the choice of priors on parameters, much more so than Bayesian estimation. This sensitivity is especially in evidence when vague priors are used for models having different numbers of parameters. In cases where alternative models involve a linear predictor, partitioning an estimated total prior variance of regression coefficients seems a reasonable expedient to dealing with this sensitivity.

All implementations of Bayesian multimodel inference involve specification of priors on parameters and priors on models; the issues that we have identified in regard to Bayes factors are present regardless of the method used to fit the models (e.g., whether one uses reversible jump Markov chain Monte Carlo [Green 1995] or implementations such as presented in our WinBUGS code.

The beauty of the Bayesian calculus is in its transparency and precision: posterior distributions are exactly determined by specification of models for data, priors for parameters, and prior model weights; there is no need for approximations of unknown precision, no need for dubious asymptotics, no need for buried assumptions. It is incumbent on the analyst to clearly articulate the reasons for choice of priors, and to evaluate the sensitivity of inferences to these assumptions. Analysts may wish to choose priors favoring more complex models, or they might wish to choose otherwise; either way, the choice should be clearly articulated.

The K-L prior, used to justify AIC weighting, might favor complex models more heavily than desired. In this case, a computationally simple, but essentially equivalent approach, is to use BIC weights, with alternative prior model weights.

Bayesian model weighting begins with a set of models, and prior probabilities that each is "Truth," given that "truth is in the model set." This formulation does not require that truth be in the model set; rather, it provides a framework for evaluating relative degrees of support in a specified context.

As emphasized by Burnham and Anderson, careful thought should go into the selection of a model set; model selection and model averaging are rather pointless exercises if none of the models under consideration is any good. But how do we know if models are any good? The standard approach (e.g., goodness-of-fit testing) is to compare the data to some prediction of what the data should look like under the model. Ultimately, our faith in a model should depend on how well it predicts future events, rather than how well it fits the data at hand. After all, it is not the elegant mathematics behind Newton's theory of gravity that convinces us that it provides a useful approximation. It is the fact that the predictions can be verified, and indeed are, in high school physics laboratories around the world. It is our impression that formal model selection and model-averaging techniques have tended to move some ecological and wildlife researchers toward the view that analysis of a single data set can be conclusive. Scientists should not rely on model weights from a single data set to form definite views on how the world works. We see the main role of such analyses as hypothesis generation and multimodel inference as an important tool in this context. However, we also believe that much more attention needs to be paid to the careful design and evaluation of evidence from follow-up studies; it is the iterative process of hypothesis generation and model evaluation that brings about advances in scientific thought. The Bayesian paradigm provides a formal mechanism for accumulating information and refining models.

We believe that ecologists and wildlife biologists ought to prefer designed experiments and to conduct such whenever possible. However, all analyses are, at some level, model-based; the purpose of designed experiments is to increase the confidence the scientist can have in the collection of models considered. Multimodel inferential techniques ought not to be regarded as an excuse for sloppy planning or data collection.

Literature Cited

Anderson, D. R., W. A. Link, D. H. Johnson, and K. P. Burnham. 2001. Suggestions for presenting the results of data analysis. Journal of Wildlife Management **65**:373–378.

Berger, J. O., and L. R. Pericchi. 1996. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association **91**:109–122.

Berger, J. O., and L. R. Pericchi. 1998. On criticisms and comparisons of default Bayes factors for model selection and hypothesis testing. ISDS Discussion Paper 97-43. Duke University, Durham, North Carolina, USA.

Box, G. E. P. 1976. Science and statistics. Journal of the American Statistical Association **71**:791–799.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods and Research **33**:261–304.

Carlin, B. P., and S. Chib. 1995. Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society (London), Series B **57**:473–484.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference (with discussion). Journal of the Royal Statistical Society (London), Series A **158**:419–466.

Claeskens, G., and N. L. Hjort. 2003. The focused information criterion. Journal of the American Statistical Association **98**: 900–916.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society (London), Series B **57**:45–97.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**:711–732.

Guthery, F. S., L. A. Brennan, M. J. Peterson, and J. J. Lusk. 2005. Information theory in wildlife science: critique and viewpoint. Journal of Wildlife Management **69**:457–465.

Hjort, N. L., and G. Claeskens. 2003. Frequentist model average estimators. Journal of the American Statistical Association, **98**:879–899.

Hoeting, J. A., D. Madigan, A. E. Rafferty, and C. T. Volinsky. 1999. Bayesian model averaging: a tutorial. Statistical Science **14**:382–417.

Kadane, J. B., and N. A. Lazar. 2004. Methods and criteria for model selection. Journal of the American Statistical Association **99**:279–290.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association **90**:773–795.

Raftery, A. E., and Y. Zheng. 2003. Discussion: inference and interpretability considerations in Frequentist model averaging and selection. Journal of the American Statistical Association **98**:931–937.

Richards, S. A. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology **86**:2805–2814.

Spiegelhalter, D., A. Thomas, and N. G. Best. 2000. WinBUGS, version 1.3 user manual. MRC Biostatistics Unit, Cambridge, UK.

Stephens, P. A., S. W. Buskirk, and G. D. Hayward. and C. Martínez del Rio. 2005. Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology **42**:4–12.

Wintle, B. A., M. A. McCarthy, C. T. Volinsky, and R. P. Kavanagh. 2003. The use of Bayesian model averaging to better represent uncertainty in ecological models. Conservation Biology **17**:1579–1590.

APPENDIX

Inconsistency of posterior mean when prior depends on sample size (*Ecological Archives* E087-159-A1). 

SUPPLEMENT

WinBUGS files for analysis of trout data (*Ecological Archives* E087-159-S1).

Ecological Archives E087-159-A1

William A. Link and Richard J. Barker. 2006. Model weights and the foundations of multimodel inference. *Ecology* 87:2626–2635.

Appendix A. Inconsistency of posterior mean when prior depends on sample size.

As an illustration of the consequences of having priors depend on sample size, consider the case of a Binomial random variable *X* consisting of *N* independent Bernoulli trials with success parameter *p*. Suppose *p* has a beta prior distribution with parameters *a* and *b*, denoted $p \sim \beta(a,b)$. The posterior distribution of *p* is $\beta(a+X, b+N-X)$ with mean

$$\frac{a+X}{a+b+N} = \left(\frac{a+b}{a+b+N}\right)\left(\frac{a}{a+b}\right) + \left(\frac{N}{a+b+N}\right)\left(\frac{X}{N}\right);$$

this is a weighted average of the prior mean $\pi = a/(a+b)$ and the maximum likelihood estimator $\hat{p} = X/N$ The important feature is that as $N \to \infty$, the weight on the prior mean goes to zero, provided that *a* and *b* are fixed.

Now suppose that $\pi = a/(a+b)$ is fixed, but that (*a* + *b*) = *kN*, for a fixed value of *k*. Then the posterior mean becomes

$$\frac{a+X}{a+b+N} = \left(\frac{k}{1+k}\right)\pi + \left(\frac{1}{1+k}\right)\hat{p} .$$

The weight on the prior mean does not go to zero as $N \to \infty$. Indeed, the posterior mean converges to something different than the MLE, hence is not consistent.

[Back to E087-159]